한국통계학회논문집
2011, 18권, 2호, 219–227

DOI: 10.5351/CKSS.2011.18.2.219

서로 다른 산포를 갖는 이변량 음이항 회귀모형에서 산포의 동일성에 대한 검정

한상문[1,a], 정병철[a]

[a]서울시립대학교 통계학과

요 약

본 연구에서는 두 반응변수의 이질적 산포를 허용하는 좀 더 일반적인 형태의 이변량 음이항 회귀모형을 삼각소거법(trivariate reduction technique)을 이용하여 제안하였다. 이 분포에서 산포의 동일성에 대한 스코어 검정과 LR 검정을 유도하고 모의실험을 통하여 각 검정법의 효율성을 비교하였다. 모의실험 결과 스코어 검정과 LR 검정 모두 명목유의수준을 제대로 유지하고 검정력도 높게 나타나 산포의 동일성을 검정하는데 효율적인 검정법으로 나타났다. 하지만 스코어 검정은 LR 검정에 비하여 계산이 간편하다는 장점이 존재하고 모의실험을 통하여 스코어 검정이 LR 검정보다 약간 나은 효율을 보였으므로 산포의 동일성에 대한 검정에서 스코어 검정의 사용을 제안하고자 한다. 더불어 실제 사례에 두 검정법을 적용하고 그 결과를 제시하였다.

주요용어: 이변량 음이항 분포, 과대산포, 스코어 검정, LR 검정.

1. 서론

이변량 포아송(Bivariate Poisson; BP)분포는 서로 상관된 이변량 계수형 자료를 모형화하는데 가장 기본적으로 사용되는 분포이다 (Holgate, 1964; Kocherlakota와 Kocherlakota, 1992, 2001). 하지만 BP분포는 두 확률변수에서 평균과 분산이 동일하다는 가정이 필요하다. 그러나 실제 문제에서 얻어지는 대부분의 자료는 분산이 평균에 비하여 크게 나타나는 과대산포(overdispersion)의 문제가 발생한다. 만일 이변량 계수자료가 과대산포가 존재하는 경우 산포를 조절할 수 있는 이변량 음이항(Bivariate Negative Binomial; BNB)분포를 사용해야 한다. Marshall과 Olkin (1990) 및 Subramaniam (1966)은 각각 두 독립적인 포아송분포와 감마분포의 혼합 및 BP분포와 감마분포의 혼합을 이용한 BNB분포를 제안하였다. 하지만 이들에 의하여 제안된 BNB분포는 모두 포아송분포와 감마분포의 혼합에 의하여 만들어지기 때문에 두 반응변수의 산포가 동일하다는 단점이 존재한다. 이와 같은 단점을 해결하기 위하여 So 등 (2011)은 일변량 음이항 분포를 따르는 3개의 확률변수를 이용하는 “삼각소거법(Trivariate Reduction Technique)”을 이용하여 두 반응변수가 서로 다른 산포를 갖을 수 있는 또 다른 형태의 BNB(Generalized BNB; GBNB)분포의 사용을 제안하였다. 사실 So 등 (2011)의 연구에 의하면 두 반응변수가 서로 다른 산포를 갖는 경우 산포의 이질성을 무시하는 모형을 사용하게 되면 회귀계수 및 회귀계수의 표준오차는 매우 편의되어 회귀계수에 대한 가설검정이 왜곡된 결과를 보여주는 것으로 나타났다.

이에 본 연구에서는 이변량 계수형 자료에 서로 다른 산포를 허용하는 GBNB 회귀모형을 고려하고자 한다. 산포의 동일성에 대한 스코어 검정통계량과 LR 검정통계량을 유도하고자 한다. 스코어 검정

이 논문은 2009년도 서울시립대학교 교내학술연구비에 의하여 연구되었음.

[1] 교신저자: (136-743) 서울 동대문구 전농동 90, 서울시립대학교 통계학과, 교수. E-mail: smhan@uos.ac.kr

은 귀무가설하에서의 최대우도추정량(MLE)만을 요구하기 때문에 LR 검정에 비하여 비교적 간단하게 계산되는 장점이 있다. 이와 같이 유도된 각 검정을 실제 사례에 적용해보고 모의실험을 통하여 각 검정의 효율성을 비교하고자 한다.

2. GBNB 회귀모형

먼저 일변량 음이항 분포는 다음과 같이 표현된다.

$$f(y) = \frac{\Gamma(\tau^{-1}+y)}{\Gamma(\tau^{-1})\Gamma(y+1)}(1+\tau\mu)^{-\frac{1}{\tau}}\left(1+\frac{1}{\tau\mu}\right)^{-y}, \quad y = 0,1,2,\ldots, \tag{2.1}$$

확률변수 Z_j가 식 (2.1)의 확률분포를 갖는 경우 $Z_j \sim \text{NB}(\mu,\tau)$라 놓도록 한다. 식 (2.1)에서 τ는 산포를 나타내는 모수로 항상 0보다 큰 값을 갖으며 $\tau = 0$인 경우 식 (2.1)은 평균이 μ인 포아송분포가 된다. 식 (2.1)의 분포를 갖는 확률변수 Z_j의 평균과 분산은 각각 다음과 같이 구해진다 (Cameron과 Trivedi, 1998).

$$E(Z_j) = \mu, \qquad V(Z_j) = \mu(1+\tau\mu). \tag{2.2}$$

이와 같은 일변량 음이항 분포를 이용하여 서로 다른 산포를 갖는 GBNB분포를 정의해보자. 먼저 Z_1, Z_2 및 Z_3를 각각 $\text{NB}(\mu_1,\tau_1)$, $\text{NB}(\mu_2,\tau_2)$ 및 $\text{NB}(\mu_3,\tau_3)$를 따르는 독립적인 확률변수라 하자. 이제 (Y_1, Y_2)를 다음과 같이 정의해보자.

$$Y_1 = Z_1 + Z_3, \qquad Y_2 = Z_2 + Z_3. \tag{2.3}$$

식 (2.3)과 같이 정의된 (Y_1, Y_2)의 결합 확률확률질량함수(probability mass function; PMF)는 다음과 같이 구해진다.

$$f(y_1,y_2) = \sum_{k=0}^{\min(y_1,y_2)} f(y_1-k)f(y_2-k)f(k) = \prod_{j=1}^{3}(1+\tau_j\mu_j)^{-\tau_j^{-1}} \sum_{k=0}^{\min(y_1,y_2)} Q(k), \tag{2.4}$$

여기서 Q_k는 다음과 같다.

$$Q(k) = \prod_{l_1=1}^{y_1-k}\frac{1+\tau_1(y_1-k-l_1)}{1+\tau_1\mu_1}\prod_{l_2=1}^{y_2-k}\frac{1+\tau_2(y_2-k-l_2)}{1+\tau_2\mu_2}\prod_{l_3=1}^{k}\frac{1+\tau_3(k-l_3)}{1+\tau_3\mu_3}\times\frac{\mu_1^{y_1-k}\mu_2^{y_2-k}\mu_3^{k}}{(y_1-k)!(y_2-k)!k!}. \tag{2.5}$$

식 (2.3)의 방법을 이용하여 이변량 이산형 확률분포를 만드는 방법을 삼각소거법(trivariate reduction technique)이라 하고 서로 상관된 두 반응변수를 생성하는 방법으로 흔히 사용되는 방법이다 (Holgate, 1964; Fayome와 Consul, 1995). 이제 식 (2.4)의 결합 PMF를 갖는 (Y_1, Y_2)에 대하여 $(Y_1, Y_2) \sim \text{GBNB}(\mu_1,\mu_2,\mu_3,\tau_1,\tau_2,\tau_3)$로 놓기로 한다. 만일 식 (2.4)에서 $\tau_1 = \tau_2 = \tau_3 = 0$가 된다면 식 (2.4)의 GBNB분포는 전통적인 BP분포로 축소될 것이다. 식 (2.4)의 확률분포를 갖는 이변량 확률변수 (Y_1, Y_2) 의 1차 및 2차 적률은 다음과 같이 구해진다.

$$\begin{aligned} E(Y_i) &= \mu_i + \mu_3, \quad i = 1,2, \\ V(Y_i) &= \mu_i(1+\tau_i\mu_i) + \mu_3(1+\tau_3\mu_3), \quad i = 1,2, \\ Cov(Y_1,Y_2) &= \mu_3(1+\tau_3\mu_3). \end{aligned} \tag{2.6}$$

식 (2.6)을 살펴보면, τ_1과 τ_2는 각각 Y_1과 Y_2에만 영향을 미치는 산포모수이며 τ_3는 두 반응변수 모두에 영향을 미치는 산포모수이다. 만일 τ_1과 τ_2가 서로 큰 차이를 보인다면 두 반응변수의 분산은 매우 큰 차이를 보일 것이다. 더불어 $\mu_3 = 0$라면 두 반응변수의 상관계수는 0이 될 것이며 식 (2.4)의 이변량 확률분포는 두 일변량 음이항 분포의 곱의 형태로 표현될 것이다. 그러므로 μ_3는 두 변수의 상관의 정도를 나타내는 상관모수로 고려할 수 있다. 즉, μ_3가 커지면 두 변수의 상관계수는 커지게 된다. 식 (2.4)에 나타난 결합 PMF는 두 반응변수의 상관과 더불어 두 반응변수에 이질적인 분산을 허용하는 좀 더 일반적인 확률분포라 할 수 있다.

이제 식 (2.4)에서 (Y_{1i}, Y_{2i}), $i = 1, \ldots, n$가 GBNB$(\mu_{1i}, \mu_{2i}, \mu_3, \tau_1, \tau_2, \tau_3)$를 따른다고 했을 때, Y_{1i}와 Y_{2i}의 기댓값이 $\mu_{1i} + \mu_3$와 $\mu_{2i} + \mu_3$임을 이용하여 다음과 같은 회귀모형을 고려해보자.

$$\mu_{1i} + \mu_3 = \exp\left(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right), \qquad \mu_{2i} + \mu_3 = \exp\left(\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2\right), \tag{2.7}$$

여기서 $\boldsymbol{x_{1i}}$와 $\boldsymbol{x_{2i}}$는 각각 $k_1 \times 1$과 $k_2 \times 1$인 설명변수 벡터를 나타내고 $\boldsymbol{\beta_1}$과 $\boldsymbol{\beta_2}$는 각각 $k_1 \times 1$과 $k_2 \times 1$인 모수 벡터를 나타낸다.

최근 들어, So 등 (2011)은 식 (2.4)의 결합 PMF를 갖는 BNB 회귀모형에서 산포의 이질성을 무시하는 경우 회귀계수 및 회귀계수의 표준오차는 심각하게 편의된다는 사실을 보였다. 그러므로 산포의 동일성 여부를 모르는 두 계수형 반응변수에 대한 모형화를 고려하는 경우, 모형적합에 앞서 두 반응변수의 산포의 동일성에 대한 검정은 반드시 실시해야만 하는 가설검정이다.

3. 산포의 동일성에 대한 검정

이제 식 (2.4)의 결합 PMF를 갖는 GBNB분포에서 다음과 같은 가설검정을 고려해보자.

$$H_0 : \tau_1 = \tau_2 \qquad vs \qquad H_1 : \tau_1 \neq \tau_2. \tag{3.1}$$

가설 (3.1)은 두 반응변수에 존재하는 산포모수의 동일성을 검정하는 가설이다. 이제 $\tau_2 = \tau_1 + \delta$로 재모수화해보자. 이와 같은 재모수화 과정을 통하면 식 (3.1)에 나타난 가설은 다음과 같이 재표현될 것이다.

$$H_0 : \delta = 0 \qquad vs \qquad H_1 : \delta \neq 0. \tag{3.2}$$

식 (3.1) 또는 (3.2)에 나타난 검정을 다루는데 있어서 가장 많이 사용되는 검정방법은 LR 검정방법(Likelihood Ratio test)과 스코어 검정을 들 수 있다. LR 검정은 제한모형(H_0)에서의 최대우도추정량(MLE)과 비제한 모형에서의 MLE를 동시에 요구하므로 계산과정이 상대적으로 복잡하게 된다. 반면 스코어 검정은 귀무가설하에서의 정보만을 요구하기 때문에 LR 검정에 비하여 계산과정이 비교적 간단하다는 장점이 있다.

가설 (3.2)에 대한 스코어 검정(Score test)과 LR 검정을 유도하기 위하여 크기가 n인 확률표본 (Y_{1i}, Y_{2i}), $i = 1, \ldots, n$으로 부터 얻은 로그우도함수는 다음과 같다.

$$\log L = \sum_{i=1}^{n}\left[-\frac{\log(1+\tau_1\mu_{1i})}{\tau_1} - \frac{\log(1+\tau_2\mu_{2i})}{\tau_2} - \frac{\log(1+\tau_3\mu_3)}{\tau_3} + \log\left(\sum_{k=0}^{\min(y_{1i},y_{2i})} Q(i,k)\right)\right], \tag{3.3}$$

여기서 $Q(i,k)$는 식 (2.5)에 나타난 $Q(k)$에서 y_1과 y_2를 각각 y_{1i}와 y_{2i}로 대치하고, μ_1과 μ_2를 μ_{1i}와 μ_{2i}로 대치하여 얻은 식을 나타낸다.

3.1. 스코어 검정

식 (3.2)에 나타난 가설에 대한 스코어 검정통계량을 유도하기 위해서는 귀무가설 하에서 얻어진 스코어 함수와 정보행렬이 필요하다. 먼저 l_i를 개별 관측값에 대한 로그우도값이라 했을 때, 귀무가설 하에서 각 모수에 대한 1차 편미분 값은 다음과 같이 얻어진다.

$$\left.\frac{\partial \log L}{\partial \boldsymbol{\beta_1}}\right|_{H_0} = \sum_{i=1}^{n} \left.\frac{\partial l_i}{\partial \boldsymbol{\beta_1}}\right|_{H_0} = -\sum_{i=1}^{n} \frac{(\hat{\mu}_{1i}+\hat{\mu}_3)\boldsymbol{x_{1i}}}{\hat{\mu}_{1i}(1+\hat{\tau}\hat{\mu}_{1i})} \left(\hat{\mu}_{1i} - \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}(y_{1i}-k)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right) = 0$$

$$\left.\frac{\partial \log L}{\partial \boldsymbol{\beta_2}}\right|_{H_0} = \sum_{i=1}^{n} \left.\frac{\partial l_i}{\partial \boldsymbol{\beta_2}}\right|_{H_0} = -\sum_{i=1}^{n} \frac{(\hat{\mu}_{2i}+\hat{\mu}_3)\boldsymbol{x_{2i}}}{\hat{\mu}_{2i}(1+\hat{\tau}\hat{\mu}_{2i})} \left(\hat{\mu}_{2i} - \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}(y_{2i}-k)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right) = 0$$

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \tau}\right|_{H_0} &= \sum_{i=1}^{n} \left.\frac{\partial l_i}{\partial \tau}\right|_{H_0} \\ &= \sum_{i=1}^{n}\left[-\frac{\hat{\mu}_{1i}}{\hat{\tau}(1+\hat{\tau}\hat{\mu}_{1i})} + \frac{\log(1+\hat{\tau}\hat{\mu}_{1i})}{\hat{\tau}^2} + \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}\left(\sum_{l_1=1}^{y_{1i}-k}\frac{y_{1i}-k-l_1}{1+\hat{\tau}(y_{1i}-k-l_1)} - \frac{(y_{1i}-k)\hat{\mu}_{1i}}{1+\hat{\tau}\hat{\mu}_{1i}}\right)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right. \\ &\quad \left. -\frac{\hat{\mu}_{2i}}{\hat{\tau}(1+\hat{\tau}\hat{\mu}_{2i})} + \frac{\log(1+\hat{\tau}\hat{\mu}_{2i})}{\hat{\tau}^2} + \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}\left(\sum_{l_2=1}^{y_{2i}-k}\frac{y_{2i}-k-l_2}{1+\hat{\tau}(y_{2i}-k-l_2)} - \frac{(y_{2i}-k)\hat{\mu}_{2i}}{1+\hat{\tau}\hat{\mu}_{2i}}\right)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right] = 0\end{aligned}$$

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \mu_3}\right|_{H_0} &= \sum_{i=1}^{n} \left.\frac{\partial l_i}{\partial \mu_3}\right|_{H_0} \\ &= \sum_{i=1}^{n}\left[\frac{1}{\hat{\mu}_{1i}(1+\hat{\tau}\hat{\mu}_{1i})}\left(\hat{\mu}_{1i} - \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}(y_{1i}-k)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right)\right. \\ &\quad \left. + \frac{1}{\hat{\mu}_{2i}(1+\hat{\tau}\hat{\mu}_{2i})}\left(\hat{\mu}_{2i} - \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}(y_{2i}-k)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right) - \frac{1}{\hat{\mu}_3(1+\hat{\tau}_3\hat{\mu}_3)}\left(\hat{\mu}_3 - \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}k\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right)\right] = 0\end{aligned}$$

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \tau_3}\right|_{H_0} &= \sum_{i=1}^{n} \left.\frac{\partial l_i}{\partial \tau_3}\right|_{H_0} \\ &= \sum_{i=1}^{n}\left[-\frac{\hat{\mu}_3}{\hat{\tau}_3(1+\hat{\tau}_3\hat{\mu}_3)} + \frac{\log(1+\hat{\tau}_3\hat{\mu}_3)}{\hat{\tau}_3^2} + \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}\left(\sum_{l_3=1}^{k}\frac{k-l_3}{1+\hat{\tau}_3(k-l_3)} - \frac{k\hat{\mu}_3}{1+\hat{\tau}_3\hat{\mu}_3}\right)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right] = 0\end{aligned}$$

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \delta}\right|_{H_0} &= \sum_{i=1}^{n} \left.\frac{\partial l_i}{\partial \delta}\right|_{H_0} = \hat{S}(\delta) \\ &= \sum_{i=1}^{n}\left[-\frac{\hat{\mu}_{2i}}{\hat{\tau}(1+\hat{\tau}\hat{\mu}_{2i})} + \frac{\log(1+\hat{\tau}\hat{\mu}_{2i})}{\hat{\tau}^2} + \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})}\left(\sum_{l_2=1}^{y_{2i}-k}\frac{y_{2i}-k-l_2}{1+\hat{\tau}(y_{2i}-k-l_2)} - \frac{(y_{2i}-k)\hat{\mu}_{2i}}{1+\hat{\tau}\hat{\mu}_{2i}}\right)\hat{Q}(i,k)}{\sum_{k=0}^{\min(y_{1i},y_{2i})}\hat{Q}(i,k)}\right], \quad (3.4)\end{aligned}$$

여기서 $\hat{\mu}_{1i} = \exp(\boldsymbol{x_{1i}}'\hat{\boldsymbol{\beta}}_{\mathbf{1}}) - \hat{\mu}_3$, $\hat{\mu}_{2i} = \exp(\boldsymbol{x_{2i}}'\hat{\boldsymbol{\beta}}_{\mathbf{2}}) - \hat{\mu}_3$를 나타내며, $\hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}, \hat{\mu}_3, \hat{\tau}$ 및 $\hat{\tau}_3$는 각각 귀무가설하에서 얻어진 $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mu_3, \tau$ 및 τ_3에 대한 ML 추정량을 나타낸다. 아울러 $\hat{Q}(i,k)$는 귀무가설 $\delta = 0(\tau_1 = \tau_2)$하에서 얻어진 $Q(i,k)$를 나타낸다.

스코어 검정통계량을 유도하기 위해서는 귀무가설 하에서 계산된 정보행렬(information matrix)가 필요하다. 하지만 본 모형에서 정보행렬은 귀무가설이 맞는다 할지라도 정확하게 계산되지 않는다. 이 경우 정보행렬의 추정치로 사용할 수 있는 한 가지 방법은 스코어 벡터의 외적(outer-product of gradient; OPG) 방법을 사용하는 것이다. 본 모형에서 $\theta = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \mu_3, \tau, \tau_3, \delta)'$라 놓는다면 OPG를 이용하

면 다음과 같은 정보행렬의 추정치를 얻을 수 있다.

$$\hat{I}(\theta) = \sum_{i=1}^{n} \frac{\partial l_i^2}{\partial\theta_j \partial\theta_k}\Big|_{H_0}, \quad j, k = 1, \ldots, k_1 + k_2 + 4 \tag{3.5}$$

식 (3.4)에 나타난 귀무가설하에서 계산된 δ에 대한 스코어 함수와 식 (3.5)에 나타난 귀무가설 하에서 계산된 정보행렬을 이용하면 가설 (3.2)를 검정하기 위한 스코어 검정통계량은 다음과 같이 얻을 수 있다.

$$T = \hat{S}(\delta)^2 J^{\delta\delta}, \tag{3.6}$$

여기서 $J^{\delta\delta}$는 정보행렬 $\hat{I}(\theta)$의 역행렬에서 δ에 대응하는 대각원소를 나타낸다. 식 (3.6)과 같이 계산된 스코어 검정통계량은 귀무가설이 맞는다는 가정 하에서 근사적으로 자유도가 1인 카이제곱분포를 따르게 된다. 그러므로 식 (3.6)에서 얻어진 검정통계량 T의 값이 자유도가 1인 카이제곱분포의 분위수 값보다 크게 되면 두 반응변수의 산포가 동일하다는 귀무가설을 기각하게 된다.

3.2. 우도비 검정

다음으로 가설 (3.2)의 산포의 동일성에 대한 우도비 검정(likelihood ratio; LR) 검정통계량을 구해보자. LR 검정은 제한모형(H_0)과 비제한 모형에서의 ML 추정량을 동시에 요구하므로 계산과정이 스코어 검정에 비하여 상대적으로 복잡하다. 본 모형에서 제한모형과 비제한 모형의 ML 추정량을 이용하면 가설 (3.2)에 대한 LR 검정통계량은 다음과 같이 구해진다.

$$\text{LRT} = -2[\log(\text{res}) - \log(\text{unres})], \tag{3.7}$$

여기서 log(res)와 log(unres)는 각각 제한모형과 비제한모형에서 얻어진 최대로그우도값을 나타낸다. 귀무가설이 맞다는 가정하에서 식 (3.7)에서 얻어진 LR 검정통계량은 근사적으로 $\chi^2(1)$분포를 따르게 된다.

4. 모의실험

앞 장에서 설명한 산포의 동일성에 대한 스코어 검정통계량 T와 LR 검정통계량 LRT의 효율성을 파악하기 위하여 모의실험을 실시하였다. 모의실험 방법은 So 등 (2011)과 유사한 방법을 사용하였다. 반응변수 (Y_{1i}, Y_{2i})의 생성은 다음과 같은 과정을 이용하였다.

Step 1. 먼저 GBNB모형에서 회귀모형은 다음과 같은 모형을 고려하였다.

$$\mu_{1i} + \mu_3 = \exp(\alpha_1 + x_{1i}\beta_1), \qquad \mu_{2i} + \mu_3 = \exp(\alpha_2 + x_{1i}\beta_2).$$

이 경우 설명변수 x_{1i}와 x_{2i}는 각각 $N(0.1/16)$에서 생성하였으며, 참 모수의 값은 $\alpha_1 = \alpha_2 = 0$으로 고정하고, $\beta_1 = \beta_2 = 1$로 고정하였다. 이와 같은 설명변수의 생성방식은 Munkin과 Trivedi (1999)의 방법과 동일한 방법이다. 이때 상관 모수인 μ_3의 값은 상관의 정도에 따른 검정통계량의 검정력을 알아보기 위하여 낮은 상관($\mu_3 = 0.1$)인 경우와 높은 상관($\mu_3 = 0.5$)인 경우를 고려하였다. 산포모수로는 $\tau_1 (= \tau)$와 τ_3의 값은 1로 고정한 상태에서 산포의 동일성에 대한 가설검정에서 추정된 유의수준과 검정력을 비교하기 위하여 $\delta (= \tau_2 - \tau)$의 값은 0에서 3까지 0.5단위로 변화시켜가며 실험하였다.

표 1: 표본수와 명목유의수준에 따른 각 검정의 추정된 유의수준과 검정력

μ_3	δ	$n = 200$				$n = 500$			
		$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.10$	
		Score(T)	LRT	Score(T)	LRT	Score(T)	LRT	Score(T)	LRT
0.1	0.0	0.061	0.057	0.118	0.107	0.060	0.058	0.120	0.118
	0.5	0.258	0.229	0.357	0.321	0.447	0.439	0.570	0.559
	1.0	0.542	0.513	0.663	0.630	0.910	0.902	0.950	0.949
	1.5	0.823	0.790	0.874	0.865	0.993	0.991	0.997	0.997
	2.0	0.938	0.925	0.964	0.964	1.000	1.000	1.000	1.000
	2.5	0.972	0.964	0.989	0.984	1.000	1.000	1.000	1.000
	3.0	0.991	0.989	0.995	0.995	1.000	1.000	1.000	1.000
0.5	0.0	0.060	0.051	0.119	0.108	0.061	0.060	0.121	0.120
	0.5	0.155	0.115	0.229	0.185	0.260	0.242	0.352	0.342
	1.0	0.303	0.255	0.407	0.358	0.632	0.622	0.736	0.712
	1.5	0.523	0.462	0.630	0.581	0.852	0.838	0.914	0.910
	2.0	0.696	0.650	0.784	0.753	0.972	0.972	0.992	0.992
	2.5	0.819	0.783	0.883	0.856	0.994	0.992	1.000	1.000
	3.0	0.891	0.853	0.935	0.925	1.000	1.000	1.000	1.000

Step 2. Step 1에서 정의된 각 모수값을 이용하여 먼저 Z_{1i}, Z_{2i}와 Z_{3i} $(i = 1, \ldots, n)$은 각각 독립적인 음이항 분포 NB(μ_{1i}, τ_1), NB(μ_{2i}, τ_2)와 NB(μ_3, τ_3)에서 생성시킨 후 $Y_{1i} = Z_{1i} + Z_{3i}$와 $Y_{2i} = Z_{2i} + Z_{3i}$의 식을 이용하여 (Y_{1i}, Y_{2i})를 생성하였다.

Step 3. Step 1과 Step 2를 통해 생성되는 (Y_{1i}, Y_{2i}) $(i = 1, \ldots, n)$를 이용하여 산포의 동일성에 대한 스코어 검정통계량과 LR 검정통계량을 계산하고 그의 유의확률을 계산하였다.

Step 4. 모든 모수조합에서 표본수 $n = 200$과 500인 경우 Step 1에서 Step 3의 과정에 대한 1,000번의 반복을 통하여 명목유의수준 $\alpha = 0.05, 0.10$인 경우 각 검정의 추정된 유의수준과 검정력을 계산하였다.

표 1은 각 표본수와 명목유의수준에 따른 각 검정의 추정된 유의수준과 검정력을 나타낸다. 표 1의 결과 중 각 칸의 맨 위에 나타난 부분이 귀무가설($H_0 : \delta = 0$ (또는 $H_0 : \tau_1 = \tau_2$))이 참인 경우이다. 먼저 명목유의수준이 0.05인 경우를 살펴보자. 이항분포에서 정규분포로의 근사를 이용하면 1000번의 반복을 통하여 추정된 유의수준이 0.036보다 작거나 0.064보다 크게 나타날 가능성은 5%미만이다. 이와 같은 기준을 통해 살펴보면 본 연구에서 고려한 스코어 검정과 LR 검정 모두 고려된 모든 실험조합에서 명목유의수준을 제대로 유지하는 것으로 나타났다. 두 검정의 검정력은 δ의 값이 0에서 멀어질수록(두 산포모수의 차이가 커질수록) 높아지고 있음을 알 수 있다. 아울러 두 검정의 검정력은 스코어 검정이 LR 검정에 비하여 약간 높은 검정력을 보이고 있다.

이상과 같은 모의실험 결과를 통하여 GBNB분포에서 산포의 동일성에 대한 검정방법으로는 스코어 검정과 LR 검정 모두 효율적인 검정방법으로 나타났다. 하지만 스코어 검정이 LR 검정에 비하여 계산이 간편하고 효율성도 떨어지지 않으므로 본 연구에서는 GBNB분포에서 산포의 동일성에 대한 검정시 스코어 검정의 사용을 제안하고자 한다.

5. 실제 자료분석

이번 장에서는 실제 자료에 본 연구에서 제안한 검정방법을 적용하고자 한다. 본 연구에서 사용한 실제자료는 Deb과 Trivedi (1997)에서 사용한 1987–1988 National Medical Expenditure Survey(NMES)

표 2: NMES자료에서 OPP(Y_1)와 HOSP(Y_2)의 이차원 분할표

		Y_2							
		0	1	2	3	4	5	6+	계
Y_1	0	2851	394	103	33	7	5	4	3397
	1	369	105	34	6	7	4	1	526
	2	149	39	8	4	2	1	1	204
	3	51	16	6	1	1	0	1	76
	4	35	6	5	1	2	0	2	51
	5	25	11	0	0	0	0	0	36
	6+	61	28	20	3	1	2	1	116
계		3541	599	176	48	20	12	10	4406

자료이다. 이 자료는 총 4406명의 미국에 거주하는 성인들의 건강상태에 관련된 6개의 반응변수와 성, 연령 및 보험가입 등 사회인구학적 특성을 나타내는 16개의 설명변수로 이루어졌다. 이 자료에 대한 자세한 설명은 Deb과 Trivedi (1997)에 자세히 나타나 있다.

본 연구에서는 이 자료에서 외과 외래진료횟수(OPP; Y_1)와 병원에 입원한 날의 수(HOSP; Y_2) 등 2변수를 반응변수로 고려하였고, 성, 연령 등 16개의 변수를 설명변수로 고려한 GBNB 회귀모형을 적용하였다. 다음 표 2는 반응변수로 고려된 두 변수에 대한 2차원 빈도표를 나타낸다.

표 2의 이차원 빈도표를 살펴보면 Y_1의 경우 표본평균과 표본분산이 각각 $\bar{Y}_1 = 0.751$과 $S_1^2 = 13.343$로 얻어지고, Y_2의 경우 $\bar{Y}_2 = 0.298$과 $S_2^2 = 0.557$로 얻어진다. 그러므로 평균과 분산의 비를 통해 Y_1의 산포가 Y_2의 산포보다 크게 나타남을 두 변수의 기초통계량을 통해서 알 수 있다. 이 자료에 산포의 동일성에 대한 스코어 검정과 LR 검정 결과 $T = 333.80\,(p = 0.0000)$로 나타나고 LR 검정통계량 값은 LRT = $354.16\,(p = 0.0000)$로 나타났다. 그러므로 유의수준을 0.05로 했을 때 이 자료에는 두 반응변수의 산포가 서로 다르게 나타나 모형적합시 반응변수의 이질적 산포를 고려한 모형을 적합해야 하는 것으로 나타났다.

6. 결론

본 연구에서는 이질적인 산포를 갖는 일반적인 이변량 음이항 회귀모형에서 산포의 동일성에 대한 스코어 검정과 LR 검정을 유도하고, 그 효율성을 모의실험을 통하여 파악하였다. 모의실험 결과 두 검정은 모두 명목유의수준을 제대로 유지하고 검정력도 높게 나타나 두 검정 모두 산포의 동일성에 대한 효율적인 검정법으로 나타났다. 하지만 본 모의실험에서 스코어 검정이 LR 검정보다 약간 높은 검정력을 보였고 계산이 간편하므로 산포의 동일성에 대한 검정에서 스코어 검정의 사용을 제안하였다.

참고 문헌

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data,* Cambridge University Press, Cambridge.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach, *Journal of Applied Econometrics*, **12**, 313–336.

Faymoe, F. and Consul, P. C. (1995). Bivariate generalized poisson distribution with some applications, *Metrika,* **42**, 127–138.

Holgate, P. (1964). Estimation for the bivariate poisson distribution, *Biometrika*, **51**, 241–245.

Kocherlakota, S. and Kocherlakota, K. (2001). Regression in the bivariate poisson distribution, *Communications in Statistics - Theory and Methods*, **30**, 815–825.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*, Marcel Dekker, New York.

Marshall, A. W. and Olkin, I. (1990). Multivariate distributions generated from mixtures of convolution and product families, In H.W. Block, A.R. Sampson and T.H. Savits(eds), *Topics in Statistical Dependence, IMS Lecture Notes - Monograph Series*, **16**, 372–393.

Munkin, M. K. and Trivedi, P. K. (1999). Simulated maximum likelihood estimation of multivariate mixed-poisson regression models with application, *Econometrics Journal*, **2**, 29–48.

So, S., Chun, H. and Jung, B. C. (2011). Bivariate negative binomial regression model with heterogeneous dispersions, *Communications in Statistics - Theory and Methods*, Submitted.

Subrahmaniam, K. (1966). A test for "Intrinsic Correlation" in the theory of accident proneness, *Journal of the Royal Statistical Society B*, **28**, 180–189.

2010년 12월 접수; 2011년 1월 채택

Tests for Equality of Dispersions in the Generalized Bivariate Negative Binomial Regression Model with Heterogeneous Dispersions

Sang Moon Han[1,a], Byoung Cheol Jung[a]

[a]Department of Statistics, University of Seoul

Abstract

In this paper, we proposed a generalized bivariate negative binomial distribution allowing heterogeneous dispersions on two dependent variables based on a trivariate reduction technique. In this model, we propose the score and LR tests for testing the equality of dispersions and compare the efficiencies of the proposed tests using a Monte Carlo study. The Monte Carlo study shows that the proposed score and LR tests prove to be an efficient test for the equality of dispersions in the view of the significance level and power. However, the score test is easier to compute than the LR test and it shows a slightly better performance than the LR test from the Monte Carlo study, we suggest the use of score tests for testing the equality of dispersions on two dependent variables. In addition, an empirical example is provided to illustrate the results.

Keywords: Bivariate negative binomial distribution, overdispersion, score test, LR test.

This work was supported by the University od Seoul 2009 Research Fund.

[1] Corresponding author: Professor, Department of Statistics, University of Seoul, Jeonnong-Dong 90, Dongdaemun-Gu, Seoul 136-743, Korea. E-mail: smhan@uos.ac.kr